Attaining the 2nd Chargaff Rule by Tandem Duplications

Siddharth Jain, **Netanel Raviv**, and **Jehoshua Bruck**
Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA
sidjain@caltech.edu, netanel.raviv@gmail.com, bruck@caltech.edu

***Abstract*—Erwin Chargaff in 1950 made an experimental observation that the count of A is equal to the count of T and the count of C is equal to the count of G in DNA. This observation played a crucial rule in the discovery of the double stranded helix structure by Watson and Crick. However, this symmetry was also observed in *single* stranded DNA. This phenomenon was termed as $2nd$ Chargaff Rule. This symmetry has been verified experimentally in genomes of several different species not only for mononucleotides but also for reverse complement pairs of larger lengths upto a small error. While the symmetry in double stranded DNA is related to base pairing, and replication mechanisms, the symmetry in a single stranded DNA is still a mystery in its function and source. In this work, we define a sequence generation model based on reverse complement tandem duplications. We show that this model generates sequences that satisfy the 2nd Chargaff Rule even when the duplication lengths are very small when compared to the length of sequences. We also provide estimates on the number of generations that are needed by this model to generate sequences that satisfy 2nd Chargaff Rule. We provide theoretical bounds on the disruption in symmetry for different values of duplication lengths under this model. Moreover, we experimentally compare the disruption in the symmetry incurred by our model with what is observed in human genome data.**

***Keywords*—reverse complement, inversion symmetry, balanced and unbalanced sequences, duplications.**

I. Introduction

Erwin Chargaff in 1950 made an experimental observation that the count of A is equal to the count of T and the count of C is equal to the count of G in DNA [4] [5]. This observation played a crucial rule in realizing the base pair grouping in DNA as discovered by Watson and Crick [6] in their double helix structure.

A similar symmetry was observed when in a long enough *single* DNA strand [18], the count of A is *almost* equal to the count to the count of T and the count of C is *almost* equal to the count of G. This symmetry was termed as 2nd Chargaff rule. This rule was verified globally for all eukaryotic chromosomes [15] as well as archael and bacterial chromosomes. However it does not hold in mitochondria, plasmids, single stranded DNA and RNA viruses.

Not only does the 2nd Chargaff rule hold for mononucleotides, but it also holds for k-mers (substrings of length k) upto length 7-8 for bacterial genomes and length 10 in human genome. There have been several papers in the past that have verified this symmetry for different values of k for more than 700 different species [1] [2] [12] [16] [17]. Given a genome of length n, the k-limit or the value of k upto which the 2nd Chargaff rule holds was empirically observed to be about $0.7 \ln n$ [19]. For human genome, the k-limit value that results from this approximation is 10. The 2nd Chargaff rule is termed as inversion symmetry (IS) in [19].

However, not being derived from any compelling principle, the existence of 2nd Chargaff rule (henceforth inversion symmetry (IS)) still remains a mystery. The presence of IS makes it plausible that most of the species share common dynamics of evolution. In [12] [19], the authors also showed that this symmetry only holds for reverse complement pairs and not for complement or any random pair of k-mers. [12] also argued that IS may be due to whole genome or segmental inverse duplications. Duplication based sequence generating models have been analysed in the past from a combinatorial [7] [9] [11] [13] [14] and probabilistic [8] [10] perspective. However, none of these duplication models analysed *reversed complement* tandem duplications. In this paper, we investigate a mathematical model for sequence generation that is based on reverse complement tandem duplications. We show that the sequences generated by this model satisfy IS after sufficiently many generations and find estimates for the number of generations required to achieve IS for different duplication lengths.

The reverse complement of a sequence $s = s_1 s_2 \cdots s_m$ is given by $s^* = s_m^c s_{m-1}^c \cdots s_2^c s_1^c$, where s_i^c denotes the complement of symbol s_i. DNA consists of 4 nucleotides or symbols A, C, G and T, where $A^c = T$, $T^c = A$, $G^c = C$ and $C^c = G$.

Example 1. *The reverse complement of $s = GTCCAGGT$ is given by $s^* = ACCTGGAC$.* □

In our model, we start from a *seed* string v and iteratively perform reverse complement tandem duplications at random positions inside v. The following example illustrates reverse complement tandem duplications:

Example 2. *Consider a seed $v = AGTTGGCA$, an instance of generating new strings by reverse complement tandem duplication process on v is*

$$\textit{Generation } 1 : v = AG\boldsymbol{TTG}GCA \rightarrow v' = AG\boldsymbol{TTG}\underline{CAA}GCA.$$

$$\textit{Generation } 2 : v' = \boldsymbol{AG}TTGCAAGCA \rightarrow v'' = \boldsymbol{AG}\underline{CT}TTGCAAGCA.$$

In generation 1, we choose a 3-length substring of v highlighted in bold and replicate its reverse complement in tandem highlighted by an underline to give v'. In generation 2, we choose a 2-length substring of v' and replicate its reverse complement to give v''. In generation 1 and generation 2 the replication or duplication length is 3 and 2 respectively. □

In this paper, we show that the reverse complement tandem duplication string system, described above in Example 2, generates strings that satisfy the 2nd Chargaff Rule or Inversion symmetry (IS) after a certain number of generations. The number of generations that are needed to attain inversion symmetry are dependent on the length of substrings that are replicated in reverse complement manner. For example, a single generation with a reverse complement tandem duplication of the entire seed is enough to satisfy the 2nd Chargaff rule (see Lemma 4). A quantity R_X^k to measure IS is defined in [3], which is based on averaging the absolute difference between the frequency of a k-mer and its reverse complement, and has been used

extensively in the literature in the past to experimentally verify the 2nd Chargaff rule for different genomes. In Figure 6, we show that the value of R_X^k computed on sequences generated by our reverse complement tandem duplication model for a suitable choice of duplication length is in consistence with the value observed in ChrX, Chr14, Chr17, Chr21 in human genome.

In section II, we provide insights as to why IS arises as a result of reverse complement tandem duplications. In section III, we formally describe our model and explain the boundary/edge effects that arise in the reverse complement tandem duplication model. We further derive upper bounds on IS disruption that is caused by the boundary effects. In section IV, we analyse our model and calculate the number of generations needed to create IS for some choices of duplication lengths. In section V, we show consistence in the R_X^k values for the sequences obtained by our model to those that are observed in different chromosomes in human genome for k-mer lengths $\leq$ 10. In section VI we conclude the paper, providing directions for future work.

II. Motivation for the Model

For any sequence Y, appending its reverse complement Y^* to itself can easily be shown to attain IS for all k-mers upto length $2|YY^*|$ (see Lemma 4).

Definition 3 *The complement of $a \in \{A, C, G, T\}$ is denoted by a^c, where $A^c = T, G^c = C, C^c = G, T^c = A$. The reverse complement of $Z \in \{A, C, G, T\}^m$ is denoted by Z^*, i.e., if*

$$Z = Z_1 Z_2 \cdots Z_m, \text{ then}$$
$$Z^* = Z_m^c Z_{m-1}^c \cdots Z_2^c Z_1^c.$$

Let u be any k-mer in Z. Let $N_Z(u)$ be the number of occurrences of u in Z, and note that

$$N_Z(u) = N_{Z^*}(u^*). \tag{1}$$

In the following lemma, let $Z \triangleq YY^*$ for some $Y \in \{A, C, G, T\}^n$.

Lemma 4 *For any k-mer u with $|u| \leq 2n$ in Z, $N_Z(u) = N_Z(u^*)$.*

Proof: For any k-mer u in Z,

$$N_Z(u) = N_Y(u) + N_{Y^*}(u) + B(u),$$
$$N_Z(u^*) = N_Y(u^*) + N_{Y^*}(u^*) + B(u^*),$$

$B(u)$ and $B(u^*)$ denote the number of times u and u^* occur at the boundary of Y and Y^* in Z, respectively. Note that from Eq. (1), $N_Y(u) = N_{Y^*}(u^*)$ and $N_{Y^*}(u) = N_Y(u^*)$, therefore in order to show $N_Z(u) = N_Z(u^*)$, we need to show

$$B(u) = B(u^*). \tag{2}$$

In order to show (2), we show for every occurrence of u on the boundary, there also exists an occurrence of u^*. Let

$$u = Y_{\max\{n-l+1,1\}} \cdots Y_n {Y_n}^c \cdots Y^c_{\max\{1,n-m+1\}}$$

where $l > 0$, $m > 0$ and $\min\{l, n\} + \min\{m, n\} = k$, then

$$u^* = Y_{\max\{1,n-m+1\}} \cdots Y_n {Y_n}^c \cdots Y^c_{\max\{1,n-l+1\}}.$$

■

It is easy to check that inversion symmetry is not guaranteed in the same way as described in Lemma 4, if $Z = YY^c$ or $Z = YY$.

Lemma 4 readily implies that the special case of a reverse complement tandem duplication of length n induces IS within 1 generation. Hence, this hints that IS, which is prevalent in many genomes, might be the result of such duplications. Since a reverse complement tandem duplication of length n is unlikely, a natural question to study in this regard is how short can reverse complement tandem duplications be in order to attain IS within a reasonable number of generations. Various aspects of this question are studied in the remainder of this paper.

III. Boundary Effect

For a sequence X and an integer k, the quantity

$$R_X^k = \frac{\frac{1}{2}\sum_{s\in\{A,C,G,T\}^k} |N_X(s) - N_X(s^*)|}{|X| - k + 1} \tag{3}$$

was defined in [3] as a means to estimate IS in X. It was also shown in [3] that R_X^k is monotone w.r.t k, i.e. $R_X^k \leq R_X^{k+1}$.

Now consider our reverse complement tandem duplication model. Let $v = xyz$, where $|y| = d$ and $|x|, |z| \geq 0$. Replicating y in a reversed complement tandem fashion results in $v_{new} = xyy^*z$. Let $y = y_1 y_2 \cdots y_d$, and $u = y_l y_{l+1} \cdots y_{l+k-1}$ be a k- length substring of y. In Lemma 4, we found that for every u in yy^*, $N_{yy^*}(u) = N_{yy^*}(u^*)$. In addition to that, due to the presence of z in v and v_{new}, we observe the following boundary effects:

1) **Boundary Effect 1:** Any k-mer that appears on the boundary of y and z in v can get lost in the creation of v_{new}, i.e., k-mers of the form $y_{d-i+1} \cdots z_j$, where $i + j = k$ and $i, j \geq 1$ may not exist in v_{new}. For example,
Example 5 . *Let $v = AGACA$ with $y = GA$ and $z = CA$, $v_{new} = AGATCCA$, the 2-mer AC exists at the boundary of y and z in v but is lost in v_{new}.* □

2) **Boundary Effect 2:** At the boundary of y^* and z in v_{new}, new k-mers are created which may not occur in v and are also not locally balanced in yy^*, i.e. k-mers of the form $y_i^c y_{i-1}^c \cdots z_1 \cdots z_j$, where $i + j = k$, and $i, j \geq 1$. For instance in Example 5, CC is a new 2-mer that is created at the boundary of y^* and z in v_{new}. Note that TC and AT are the other newly created 2-mers in v_{new} but they lie entirely in $yy^* = GATC$. They are locally balanced in yy^* by their reverse complements GA and AT respectively.

Definition 6 *We define a recursive process of generating strings by reverse complement tandem duplication as follows:*

- **Seed:** $v_0 = x_0 y_0 z_0$
- **Replication operation** ($\mathcal{T}_{R_c}$) : $\mathcal{T}_{R_c}(v_i) = v_{i+1} = x_i y_i y_i^* z_i$, $|x_i|, |z_i| \geq 0$ *and* $|y_i| = d_i$, $d_i > 0$, $\forall i \geq 0$.
- $\mathcal{T}_{R_c}^m(v_j) = \mathcal{T}_{R_c}^{m-1}(v_{j+1})$ $\forall m > 0$ *and* $\mathcal{T}_{R_c}^0(v_j) = v_j$.

Note that y_i is chosen uniformly at random, and let $X = \mathcal{T}_{R_c}^g(v_0)$, where g is the number of generations that we wish to study. Let $X = \mathcal{T}_{R_c}^g(v_0)$.

In order to do a cleaner analysis of R_X^k for X generated by reverse complement tandem duplication system $\mathcal{T}_{R_c}$ above, we wish to upper bound the Boundary effect 1 and Boundary effect 2. We do so by computing $\Delta_{(k,g)}$, which measures the worst case impact of Boundary effect 1 and Boundary effect 2 on R_X^k after g generations.

In each operation $\mathcal{T}_{R_c}$, we lose $k-1$ k-mers due to Boundary Effect 1 and gain $k-1$ k-mers due to Boundary Effect 2. Therefore, in each generation the numerator of R_X^k has a worst case change of $2(k-1)$. Hence, after g operations the worst case effect on numerator of R_X^k due to boundary effects is $2(k-1)g$. Also note that $|X| = |v_0| + \sum_{i=0}^{g-1} d_i$, and hence $\Delta_{(k,g)}$ is given by

$$\Delta_{(k,g)} = \frac{2(k-1)g}{|v_0| + \sum_{i=0}^{g-1} d_i - k + 1}. \tag{4}$$

Therefore after g generations, R_X^k is given by $|R_X^k - Q_X^k| \leq \Delta_{(k,g)}$, where Q_X^k is the approximation of R_X^k by ignoring the boundary effects 1 and 2 after g generations.

It is notable here that in $\mathcal{T}_{R_c}$ (Definition 6), if any string y_i of length d_i is duplicated to create $y_i y_i^*$, then by Lemma 4, it balances the occurence of any k-mer for all $k \leq 2d_i$ in $y_i y_i^*$. If every symbol in v_0 is chosen at some stage in the generation process as a substring of y_i, then once all the symbols in v_0 have been chosen, Q_X^k will be 0 for all $k \leq 2\min\{d_i\}_{i=1}^{g}$. In the next section, we find for a given $\epsilon > 0$, number of generations g that are needed to obtain $Q_X^k \leq \epsilon$ for different choices of duplication lengths; and consequently, the number of generations g which is required to obtain $R_X^k \leq \Delta_{(k,g)} + \epsilon$ for a given k. Intuitively one can expect lesser number of generations for a higher value of duplication length. We define these ideas more formally in the following section.

IV. Results and Discussions

A. Balanced and Unbalanced Segments

Definition 7 *Consider the string replication system $\mathcal{T}_{R_c}$ given in Definition 6. A symbol $a \in v_i$ is called* balanced *if it belonged to some y_j or y_j^* for $j < i$, and otherwise it is* unbalanced.

Note that all symbols in $Z = YY^*$ given in Lemma 4 are balanced.

Definition 8 *Let u be a substring of v_i. Let a and b be the symbols preceeding and succeeding u in v_i respectively. u is called a balanced segment of v_i if all the symbols in u are balanced and a, b are unbalanced.*

Let u be a substring of v_i. Let a and b be the symbols preceeding and succeeding u in v_i respectively. u is called an unbalanced segment of v_i if all the symbols in u are unbalanced and a, b are balanced.

Note that in the case where u is a prefix or suffix of v_i, we ignore symbol a and b in Definition 8 accordingly.

Note that all the symbols in v_0 are unbalanced, hence v_0 is an unbalanced segment. We will now investigate the generation of balanced segments in v_i for $i \geq 1$ for the string replication system $\mathcal{T}_{R_c}$ described in Definition 6.

B. Generation of Balanced segments

In every operation $\mathcal{T}_{R_c}(v_i)$ for $i \geq 0$, either a new balanced segment is added or some previously existing balanced segment(s) is modified. The operation $\mathcal{T}_{R_c}(v_i)$ uniformly and randomly chooses a substring y_i of length d_i in v_i and replicates it to give $v_{i+1} = x_i y_i y_i^* z_i$. The addition and modification of balanced segments is described below.

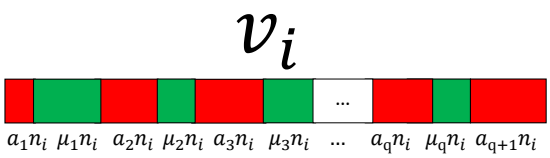

Fig. 1. Red and green segments represent unbalanced and balanced segments in v_i respectively. Each $a_j n_i (1 \leq j \leq q_i + 1)$ represents a unbalanced segment. Each $\mu_j n_i (1 \leq j \leq q_i)$ represents a balanced segment. Note $q_i \leq i$.

1) *Addition:* If all the symbols in y_i are unbalanced and the symbols before and after y_i are both unbalanced in v_i, then $y_i y_i^*$ is *added* as a new balanced segment in v_{i+1}, thereby increasing the number of balanced segments in v_{i+1} by 1.
2) *Modification:* If some of the symbols in y_i are balanced or y_i is preceeded/succeeded by a balanced symbol in v_i, then $y_i y_i^*$ *modifies* previously created balanced segment(s), thereby not increasing the count of balanced segments in v_{i+1}.

Addition and modification operations are described in Figure 2 and Figure 3 respectively. It is clear from the description of *addition* and *modification* above, that v_i has at most i balanced segments.

Figure 1 shows *balanced* (green) and *unbalanced* (red) segments in v_i. Here μ_j and a_j represent the fraction of

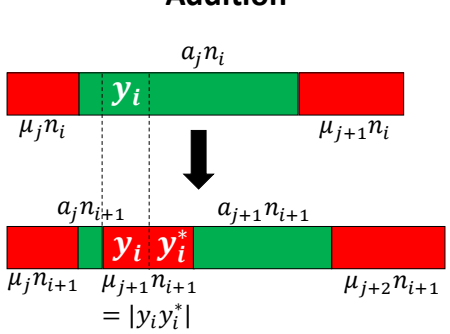

Fig. 2. Addition Operation: All the symbols of y_i are unbalanced in v_i and the symbol before and after y_i are also unbalanced in v_i. $y_i y_i^*$ is added as a new balanced segment in v_{i+1}. Note that the count of balanced segments is 1 more than the count of balanced segments in v_i.

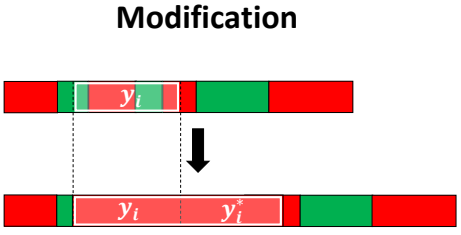

Fig. 3. Modification Operation: Some symbols of y_i are balanced in v_i. Hence $y_i y_i^*$ in v_{i+1} is a modification of the previously created balanced segments in v_i. Note that the count of balanced segments in v_{i+1} has not increased from v_i. Infact in this particular instance, it has decreased.

v_i covered by the j-th balanced and unbalanced segments respectively. Therefore,

$$n_i = \sum_{j=1}^{q_i} \mu_j n_i + \sum_{j=1}^{q_i+1} a_j n_i. \tag{5}$$

Let N_i denote the total length of balanced segments in v_i, i.e. $N_i = \sum_{j=1}^{q_i} \mu_j n_i$. We also note that

$$n_i - N_i = \sum_{j=1}^{q_i+1} a_j n_i. \quad (6)$$

Let X_i and Y_i denote the total length of unbalanced and balanced segments in y_i, respectively, and note that N_i, X_i and Y_i are random variables that satisfy

$$\begin{aligned} X_i + Y_i &= |y_i| = d_i \\ N_{i+1} &= N_i + 2X_i + Y_i \\ &= N_i + X_i + d_i. \end{aligned}$$

In turn, this readily implies that

$$E[N_{i+1}|N_i, a_1, a_2, \cdots, a_{q_i+1}] = N_i + d_i + E[X_i|N_i, a_1, a_2, \cdots a_{q_i+1}]. \quad (7)$$

We compute $E[X_i|N_i, a_1, a_2, \cdots a_{q_i+1}]$ for $d_i = d$ for all i and some $d > 0$.

Lemma 9 *Let $d_i = d$. Then $\forall\ i \geq 0$, the length of each balanced segment in v_i is at least $2d$.*

Proof: Observe that for any i, any *newly added* balanced segment in v_i, i.e., one that was generated by the most recent application of $\mathcal{T}_{R_c}$, is of length at least $2d$. ■

We derive $E[X_i|N_i, a_1, a_2, \cdots a_{q_i+1}]$ by using

$$E[X_i|N_i, a_1, a_2, \cdots a_{q_i+1}] = \Sigma_{l=1}^{d} P(X_i \geq l|N_i, a_1, a_2, \cdots a_{q_i+1}). \quad (8)$$

Lemma 10 *For $d_i = d\ \forall i$, y_i can overlap with at most 2 balanced segments in v_i.*

Proof: From Lemma 9, the length of each balanced segment in v_i is at least $2d$. We have the following 4 cases:

- *Case 1:* y_i does not overlap with any balanced segment in v_i.
- *Case 2:* either some prefix of y_i overlaps with a suffix of a balanced segment j for some j in v_i or some suffix of y_i overlaps with a prefix of a balanced segment j for some j in v_i but not both.
- *Case 3:* y_i is a substring of some balanced segment j.
- *Case 4:* some prefix of y_i overlaps with a suffix of a balanced segment j and some suffix of y_i overlaps with a prefix of next balanced segment after j in v_i for some j.

In case 2 and 3 above, y_i overlaps with 1 balanced segment and in case 4 y_i overlaps with 2 balanced segments. ■

We derive $E[X_i|N_i, a_1, a_2, \cdots a_{q_i+1}]$ by using

$$E[X_i|N_i, a_1, a_2, \cdots a_{q_i+1}] = \Sigma_{l=1}^{d} P(X_i \geq l|N_i, a_1, a_2, \cdots a_{q_i+1}). \quad (9)$$

Using Lemma 10,

$$P(X_i \geq l|N_i, a_1, a_2, \cdots, a_{q_i+1}) = \sum_{j=2}^{q_i} I(a_j n_i \geq l)\frac{a_j n_i + d - 2l + 1}{n_i - d + 1} + I(a_1 n_i \geq l)\frac{a_1 n_i - l + 1}{n_i - d + 1} + I(a_{q_i+1} n_i \geq l)\frac{a_{q_i+1} n_i - l + 1}{n_i - d + 1} \quad (10)$$

$I(.)$ represents the indicator function.
Solving (9) and (10) and using (6) gives

$$E[X_i|N_i, a_1, a_2, \cdots, a_{q_i+1}] = \frac{(n_i - N_i - a_1 n_i - a_{q_i+1} n_i)d}{n_i - d + 1} + \sum_{l=1}^{\min(a_1 n_i, d)} \frac{a_1 n_i - l + 1}{n_i - d + 1} + \sum_{l=1}^{\min(a_{q_i+1} n_i, d)} \frac{a_{q_i+1} n_i - l + 1}{n_i - d + 1}.$$

To do numerical simulations for the recursive Eq. (7), we can omit a_1 and a_{q_i+1} by approximating $E[X_i|N_i, a_1, a_2, \cdots, a_{q_i+1}]$. We do this by stitching the end of v_i with its start, thereby making v_i circular. Let X_i' denote the number of unbalanced symbols chosen in this circular version of v_i. In turn, $P(X_i' \geq l|N_i, a_1, a_2, \cdots, a_{q_i+1})$ is given by

$$P(X_i' \geq l|N_i, a_1, a_2, \cdots, a_{q_i+1}) = \sum_{j=2}^{q_i} I(a_j n_i \geq l)\frac{a_j n_i + d - 2l + 1}{n_i} + I((a_1 + a_{q_i+1}) n_i \geq l)\frac{(a_1 + a_{q_i+1}) n_i + d - 2l + 1}{n_i} \quad (11)$$

Solving (9) and (11) and using (6) gives

$$E[X_i'|N_i, a_1, a_2, \cdots, a_{q_i+1}] = \frac{(n_i - N_i)d}{n_i} = E[X_i'|N_i]$$

We can now use $E[X_i'|N_i, a_1, a_2, \cdots, a_{q_i+1}]$ as an approximation for $E[X_i|N_i, a_1, a_2, \cdots, a_{q_i+1}]$. Using (7) we get,

$$E[N_{i+1}|N_i] \approx N_i + d + \frac{(n_i - N_i)d}{n_i}. \quad (12)$$

Figure 4 is obtained by using $E[N_i|N_{i-1}]$ as approximation for N_i and $N_0 = 0$ in Eq. (12). It shows the number of generations g needed such that $\frac{N_g}{n_g} = 1 - \epsilon$ for $\epsilon = 0.00005$, $|v_0| = 10000$ for different values of d.

When $d_i = d\ \forall i$, $\Delta_{(k,g)}$ using Eq. (4) is given by

$$\Delta_{(k,g)} = \frac{2(k-1)g}{|v| + gd - k + 1}$$

We also see that

$$\lim_{g \to \infty} \Delta_{(k,g)} = \frac{2(k-1)}{d}. \quad (13)$$

Further since $\Delta_{(k,g)}$ is an increasing function in g for $|v| > k - 1$, we have for a given k and d that $\Delta_{(k,g)} \leq \frac{2(k-1)}{d}$, when $|v| > k - 1$. Figure 5 shows the variation of $\Delta_{(10,g)}$ with g for different values of d at $|v| = 10000$. We see from

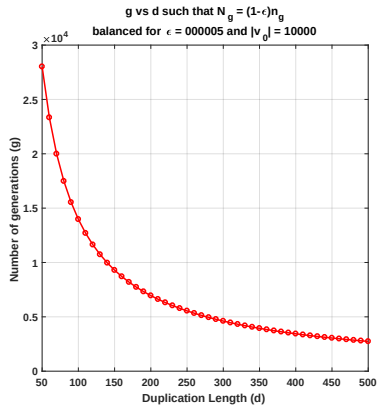

Fig. 4. Number of generations (g) needed such that $\frac{N_g}{n_g} = (1-\epsilon)$ for $\epsilon = 0.00005$, $|v_0| = 10000$ and $d_i = d\ \forall i \geq 0$.

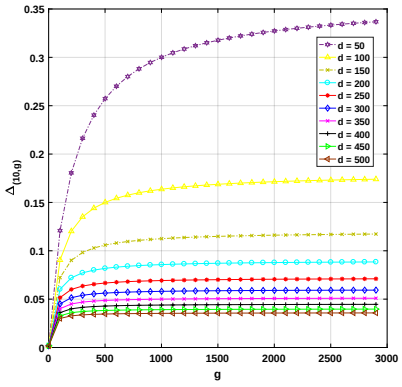

Fig. 5. Variation of $\Delta_{(10,g)}$ vs g for d = 50, 100, 150, 200, 250, 300, 350, 400, 450, 500. and $|v| = 10000$. We can see that a larger value of d diminishes the boundary effects.

Figure 5 that $\Delta_{(10,g)} \approx 0.09$ for $d = 200$ for $g > 1000$ generations. Therefore $|R_X^{10} - Q_X^{10}| \leq 0.09$ for $d = 200$ after 1000 generations. Further, from Figure 4 we observe that after about 7000 generations at least $(1 - 0.00005)$ fraction of the sequence is balanced, which implies $Q_X^{10} \leq 0.00005$, and hence $R_X^{10} \leq 0.09005$ for $d = 200$ after 7000 generations using the reverse complement tandem duplication model. Similar bounds on R_X^{10} can be derived for other values of d using Figures 2 and 3. Note that the choice of $\epsilon = 0.00005$ for generating the plot in Figure 4 is arbitrary here and similar plots can be obtained for other values of ϵ by using (12). Moreover, similar bounds can be obtained for other values of k by calculating $\Delta_{(k,g)}$.

C. Unbalanced Shorter Sequences

Inversion symmetry however is not observed in the shorter segments of the genome [19]. For example, if a segment of length 5000 is selected from our genome, it will not possess inversion symmetry upto $k = 10$. Our generative model based on reverse complement tandem duplication is also in consensus with this experimental observation. More precisely, as more and more duplication happens, the k-mers that became balanced in the creation of yy^* will pull apart in future generations if duplication happens somewhere inside yy^*. This distance between a k-mer and its reverse complement arises due to extra duplications that happen in the segments that lie in between them. As a result, the whole sequence remains balanced however the shorter segments inside it become unbalanced. This is illustrated using the following example:

Example 11. *Consider*

$$v = GTCCG\underline{AGCACTGA}AGTCA.$$

Let y denote the underlined substring of v. u is obtained by duplicating y in v in reversed complement tandem. Therefore

$$u = GTCCG\underline{AGCACTGA}\boldsymbol{TCAGTGCT}AGTCA.$$

y^* is denoted by the bold portion in u. Let us now focus on the 2-mer CA and its reverse complement TG in y and y^* respectively. We note that $|y| = 8$. Below we highlight this 2-mer and its reverse complement in yy^* in u.

$$u = GTCCGAG\boldsymbol{CA}CTGATCAG\boldsymbol{TG}CTAGTCA.$$

Now if we further duplicate the underlined portion in

$$u = GTCCGAG\boldsymbol{CA}\underline{CTGATCAG}\boldsymbol{TG}CTAGTCA$$

to get

$$u' = GTCCGAG\boldsymbol{CA}\underline{CTGATCAGCTGATCAG}\boldsymbol{TG}CTAGTCA$$

We observe that in u CA and TG were apart by 8 symbols in u and by 16 symbols in u'. More such duplications in between CA and TG in the future generations will pull them further apart. Note that the duplication length chosen here cannot be more than 8 as the distance between CA and TG is 8. □

We analyze the phenomenon explained in Example 11 above by defining Δ_0 as the initial distance between a k-mer and its reverse complement when they are created, and Δ_i as their distance after i generations. The expected behavior of Δ_i can be modeled by the equation below:

$$E[\Delta_{i+1}|\Delta_i] = \Delta_i(1 + \frac{d}{n_i - d + 1}). \tag{14}$$

For $n_i \gg d$, $E[\Delta_{i+1}|\Delta_i] \approx \Delta_i(1 + \frac{d}{n_i})$. Note $n_{i+1} = n_i + d$. Therefore by approximating Δ_i with $E[\Delta_i|\Delta_{i-1}]$, we have

$$E[\Delta_m|\Delta_0] \approx \Delta_0(1 + \frac{md}{n_0}). \tag{15}$$

Note here $d \leq \Delta_0 \leq 2d - k - 1$.

Example 12. *For $k = 10$, $n_0 = 10000$ and $d = 200$. Using (15) above, we have $E[\Delta_m|\Delta_0] \approx \Delta_0(1 + 0.02m)$. Now using $200 \leq \Delta_0 \leq 389$, we have $200 + 4m \leq E[\Delta_m|\Delta_0] \leq 389 + 7.78m$. We see that $200 + 4m = 5000$, for $m = 1200$, which implies that after 1200 generations $E[\Delta_{1200}|\Delta_0] \geq 5000$. In Figure 4, we see that for $d = 200$, $\frac{N_g}{n_g} \geq 0.99995$ only after about 7000 generations which implies that the balanced k-mers would have been pulled further apart and will not be localized in smaller blocks of length 5000 inside the sequence.* □

V. Experimental Findings

In Figure 6, we compare the experimentally observed value of R_X^k for different sequences X and $1 \leq k \leq 10$. The sequences chosen are chromosomes $X, 14, 17, 21$ (shown by solid lines) in the human genome (Hg38), sequences generated by tandem reverse complement duplication system discussed in the paper for $d = 20, 50, 200, 500$ (shown by dashed lines). We observe that tandem at $d = 200$ is well in consistence with ChrX and Chr14. We also observe that $d = 50$ tandem is in consistence with Chr17 and Chr21 for $k = 9, 10$. We have further added 2 more plots (shown by dotted lines) that

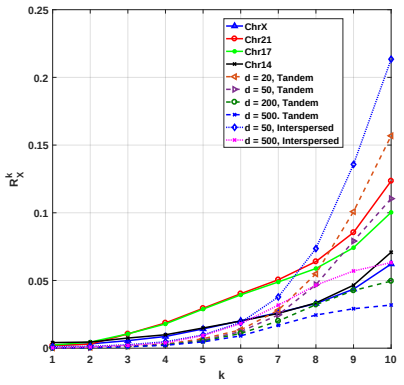

Fig. 6. Comparison of R^k_X value calculated experimentally in chromsomes X, 14, 17 and 21 (solid lines) with those obtained by reverse complement tandem (dashed lines) and interspersed (dotted lines) duplications for different duplication lengths d.

model reverse complement duplications done in an interspersed manner. In interspersed duplication, unlike tandem the chosen substring is replicated at any location and not necessarily next to the original string. An example illustrating interspersed duplication is given below

Example 13. *Consider a seed $v = AGTTGGCA$, an instance of generating new strings by reverse complement interspersed duplication process on v is*

$$\textit{Generation } 1 : v = AG\boldsymbol{TTG}GCA \rightarrow v' = AG\boldsymbol{TTG}GC\underline{CAA}A.$$

$$\textit{Generation } 2 : v' = \boldsymbol{AG}TTGGCCAAA \rightarrow v'' = \boldsymbol{AG}TTG\underline{CT}GCCAAA.$$

In generation 1, we choose a 3-length substring of v highlighted in bold and replicate its reverse complement in an interspersed manner highlighted by an underline to give v'. In generation 2, we choose a 2-length substring of v' and replicate its reverse complement to give v''. In generation 1 and generation 2 the replication or duplication length is 3 and 2 respectively. □

In the plot in Figure 6, we have included two plots where the sequences are generated by interspersed duplication and the duplication length is 50 and 500. The site where the reverse complement duplicate is placed is chosen uniformly and randomly in the interspersed model. We see that at d = 500, interspersed duplication is in consistence with values observed in ChrX and Chr14 for $k \leq 7$.

These plots suggest that reverse complement duplications can potentially be playing a key role in the evolution of genome. We believe that the inconsistencies with chromosome data that are seen for some $k's$ can be attributed to point mutations which was not taken into account.

VI. Conclusion

We showed that the reverse complement tandem duplication model generates sequences satisfying 2nd Chargaff Rule. Moreover, even when the length of duplication is chosen to be a constant d which is very small as compared to the sequence length, this symmetry can be obtained using our model. Further, we provided estimates on the number of generations needed to create this symmetry given a choice of duplication length(s). In our analysis, we found an upper bound given in (4) on the disruption caused by boundary effects.

We see that the error due to boundary effect given in Eq. 13 for $d = 50$ and $k = 10$ is 0.36. However, we note from Figure 6, the R^{10}_X value obtained experimentally when we generate sequences from our model for $d = 50$ is 0.11. This means that the theoretical bound on R^k_X given by $R^k_X \leq Q^k_X + \Delta_{(k,g)}$ obtained in this paper is loose for lower values of duplication length d ($d < 150$). For such lower values of d, a finer boundary effect analysis is needed and is deferred to future work. Another interesting question is how does the value of d affect the probability distribution of k-mers observed using this model of sequence generation and comparing it with the k-mer distribution observed in real DNA.

Acknowledgements

This work was supported in part by the NSF Expeditions in Computing Program - The Molecular Programming Project. The work of Netanel Raviv was supported in part by the postdoctoral fellowship of the Center for the Mathematics of Information (CMI), Caltech, and in part by the Lester-Deutsch postdoctoral fellowship.

References

[1] V. Afreixo, C.A.C. Bastos, S.P. Garcia, J.M.O.S. Rodrigues, A.J. Pinho, P.J.S.G. Ferreira, *"The breakdown of the word symmetry in the human genome,"* J Theor Biol. 2013;335:1539.

[2] V. Afreixo, J.M.O.S. Rodrigues, C.A.C. Bastos, *"Analysis of single-strand exceptional word symmetry in the human genome: new measures,"* Biostatistics. 2015;16(2):20921

[3] P-F. Baisnee, S. Hampson , P. Baldi, *"Why are reverseary DNA strands symmetric?,"* Bioinformatics. 2002;18:102133.

[4] E. Chargaff, *"Chemical specificity of nucleic acids and mechanism of their enzymatic degradation,"* Experientia. 1950;6(6):2019

[5] E. Chargaff, *" Structure and function of nucleic acids as cell constituents,"* Federal Proc. 1951;10:6549.

[6] F. Crick, J. D. Watson, *"Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid,"* Nature, 1953;, 171:737-8.

[7] J. Dassow, V. Mitrana, and A. Salomaa, *"Operations and language generating devices suggested by the genome evolution,"* Theoretical Computer Science, vol. 270, no.1 , pp. 701-738, 2002

[8] O. Elishco, F. Farnoud, M. Schwartz, J. Bruck, *"The capacity of some Polya string models,"* Proceedings of IEEE International Symposium on Information Theory (ISIT), 2016, pp. 270-274.

[9] F. Farnoud, M. Schwartz, J. Bruck, *"The capacity of string-duplication systems,"* IEEE Trans. Inf. Theory, vol. 62, no. 2, pp. 811-824, Feb. 2016.

[10] F. Farnoud, M. Schwartz, J. Bruck, *"A stochastic model for genomic interspersed duplication,"* Proceedings of IEEE International Symposium on Information Theory (ISIT), 2015, pp. 904-908.

[11] S. Jain, F. Farnoud, J. Bruck, *"Capacity and Expressiveness of genomic tandem duplication,"* IEEE Trans. Inf. Theory, vol. 63, no. 10, pp. 6129-6138, Oct. 2017.

[12] S-G. Kong, W-L. Fan, H-D. Chen, Z-T Hsu, N. Zhou, B. Zheng, H-C. Lee, *"Inverse symmetry in complete genomes and whole-genome inverse duplication,"*PlosOne. 2009;4:e7553.

[13] P. Leupold, C. Martin-Vide, and V. Mitrana, *"Uniformly bounded duplication languages,"* Discrete Applied Mathematics, vol. 146, no. 3, pp. 301-310, 2005.

[14] P. Leupold, V. Mitrana, and J. M. Sempere, *"Formal languages arising from gene repeated duplication,"* in Aspects of Molecular Computing, Springer, 2004, pp. 297-308.

[15] D. Mitchell, R. Bridge, *"A test of Chargaff's secong rule,"*Biochem Biophys Res Commun. 2006;340(1):90-4.

[16] B.R. Powdel, S.S. Satapathy, A. Kumar, P.K. Jha, A.K. Buragohan, M. Borah, S.K. Ray, *"A Study in Entire Chromosomes of Violations of the Intra-strand Parity of Complementary Nucleotides (Chargaffs Second Parity Rule),"* DNA Res. 2009;16:32543.

[17] D. Qi, A.J. Cuticchia,*"Compositional symmetries in complete genomes,"* Bioinformatics. 2001;17:557-9.

[18] R. Rudner, J. D. Karkas, E. Chargaff, *"Separation of B. subtilis DNA into complementary strands. III. Direct Analysis,"* Proc. National Acad Sci USA 1968;60:921-2.

[19] S. Shporer, B. Chor, S. Rosset, D. Horn, *"Inversion symmetry of DNA k-mer counts: validity and deviations,"* BMC Genomics. 2016, 17 (1): 696.